## Anaphora, Coreference Resolution, Language Identification

### Accomplished Tasks

* …

### Checklist

* …

## Apache Lucene

### Accomplished Tasks

* Generate collection frequency, document frequency, inverted document frequency, and term frequency in the Gutenberg Minicorpus. *(This data should be accumulated using Apache Lucene 7.3.0, but may also be gathered for a modified programming assignment #3 for Information Retrieval.)*
  + This was accomplished by using PostingsEnum and TermsEnum for each of the documents in the corpus, but also required creation of a custom FieldType to store term vectors in the generated index.
* Generate Heaps’ Law coordinate data for the documents of the Gutenberg Minicorpus, and use those coordinate points to generate figures for verifying Heaps’ Law on the corpus.
  + This was accomplished by using TokenStream and CharTermAttribute for iterating through all of the tokens of corpus entry document content, using a custom Heaper class to generate the data. The result – CSV files – was then piped through a custom Python script to generate figures.
* …

### Checklist

* **Doc:** Notes
  + Locate execution commands for adding corpa to ElasticSearch
  + Utilize the *IR: Libraries and Frameworks* paper to look into ElasticSearch, Indri, Solr
  + **Resources:**
    - *Software:* ElasticSearch w/ Kibana; Indri; Solr; srcML; Luke/LukeFX
    - *Corpa:* Reuter’s; (Unknown corpus)
    - *Books and Text:* Lucene 4 Cookbook; Lucene in Action, Second Edition; Mastering ElasticSearch, Second Edition; and ElasticSearch Tutorial (TutorialsPoint); IR Textbook
  + Restructure how generated resources are listed. This should probably be in its **own** document
* **Doc:** Apache Lucene API
  + *Shift the API document (or attempt to) to LaTeX*
  + Implementation research
    - (Intermediate) Look into TopFieldDocs search(…) and research ‘Sort’ for IndexSearcher
    - (Intermediate) Look into DateTools in the document package
    - Expert implementation: *custom analyzers, codecs, similarities*
  + Document Modifications and Child Resources
    - Generate charts/diagrams for the core packages *(not inside of the document though!)*
* Look at the various analyzers of Lucene for pre-text processing (not including foreign language analyzers)
* Implement methodology to verify Zipf’s law on the Gutenberg corpus
* **Programming**
  + **Proj:** Lucene API implementation *(Java)*
    - Program doesn’t like to print terms if documents have no content for .W
    - Implement a single document toggle (both file and string)
    - Implement check for similarity on indexes (if a method doesn’t already exist)
    - Implement a toggle for switching similarity/analyzer
  + **Proj:** Plotter *(Python)*
  + Write a Java file to parse Cranfield Collection format into JSON objects for ElasticSearch
  + *The folder ‘out’ is reserved for program compilation; program output goes to the ‘prog’ folder*

The following notes have been made for questioning:

* Gudivada’s notes from the e-mail on 05/19 specify Lucene has a **NumericRangeQuery**; however, this class couldn’t be located in the Lucene API for 7.3.0. It does exist in the version 4 API. In the [version 6 API](https://lucene.apache.org/core/6_0_0/core/index.html), NumericRangeQuery has become **LegacyNumericRangeQuery**. *(It seems as though NumericRangeQuery was phased out of the API, and its features implemented in the* ***IntPoint****, etc. classes.)*
* The Apache Lucene 7.3.0 API overview specifies the following code line: “*ScoreDoc[] hits = isearcher.search(query, null, 1000).scoreDocs;*”. There is a problem with this line, where the parameters of the search(…) function are not suitable. The API lists the parameters as being (Query, int, Sort), not (Query, Sort, int). *(It seems as though this might be a typographical error.)*

### Tables

|  |  |
| --- | --- |
| Function Name | Description |
| … | *…* |
| … | *…* |

|  |  |
| --- | --- |
| *Class* |  |
| Function Name | **Description** |
| … | *…* |
| … | *…* |

|  |  |  |
| --- | --- | --- |
| ***Jar file*** |  |  |
| **package** | | |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| ***Document Title*** | ***Date*** | ***Author*** |
| Description  **Location:** …  **Generated Children Resources:**   * *Resource*, Date   + **Location:** …   + Description of resource | | |